A Semiparametirc Approach for Canonical Calibration in Multi-Label Classification Tasks

Arkapal Panda, Utpal Garain

Computer Vision and Pattern Recognition Unit
Indian Statistical Institute, Kolkata

Abstract

Existing methods for Estimation of Expected Calibration Error (ECE) in Multi-Label Classification (MLC) problems often overlook the interdependencies among labels which effects the estimate negatively. As a remedy, we propose an unbiased, differentiable, trainable estimator of ECE using Copula, addressing canonical calibration beyond marginal calibration. Our method leverages the kernel trick to model continuous distributions from discrete labels and uses a semi-parametric approach: non-parametric marginals and parametric Copula to model the calibration error. We theoretically prove that our estimator is unbiased and converge to true ECE. Also, the estimator is consistent, differentiable and trainable in terms of L_p calibration error. Finally, we use our estimator in conjuction with conventional loss functions for calibration regularized risk minimization.

Importance of Callibration in MLC tasks

- In medical image analysis, a X-Ray or CT-scan image may have multiple indications related to different diseases.
- In several object recognition tasks including autonomous driving generally have multiple labels associated at once.
- In e-commerce, products often belong to multiple categories and when recommending products, users might have diverse interests.

In all high stake tasks, calibration is needed so the the model is less prone to make overconfident mistakes [4]. A well calibrated model properly quantify the uncertainty regarding an outcome so that when the confidence of a prediction is relatively low, the control can be shifted to humans.

Expected Calibration Error in MLC tasks

Let $f : \mathcal{X} \to [0,1]^K$, be Neural Network for multi-label classification. We measure the the (mis-)calibration of the model in terms of L_p calibration error, defined below.

$$CE_p(f) = \left(E\left[\left|\left| E[\mathbf{y}|f(\mathbf{x})] - h(\mathbf{x}) \right|\right|_p^p \right] \right)^{\frac{1}{p}}$$

This is also known as Expected Calibration Error(ECE) [5] for $p = 1$. Note that, as $f = (f_1, \ldots, f_K)$ and $\mathbf{y}$ is multi-hot encoded vector of dimension K (number of labels), the conditional expectation in the right hand side is a vector of dimension K.

Problem Statement

- It has been observed that the labels in multi-label datasets are often interdependent. To properly estimate the Calibration Error we need to incorporate the information of label dependencies.
- Previous works [2, 3] have suggested some methods of calibrating MLC tasks but they also didn't consider label dependencies while estimating ECE.
- We have proposed novel estimator of calibration error for MLC tasks which takes label dependency into account.
- Theoretically we have demonstrated the effectiveness and validity our ECE estimator by showing it is consistent, unbiased and differentiable. Consistency and differentiability can also be proved for $CE_p(f)$
- We use our estimator as an auxiliary loss in conjunction with the conventional loss functions for calibrated regularised loss minimization

Proposed Method

- We require that this estimator is consistent and differentiable, such that we can train it in a calibration error regularized risk minimization framework. The estimator is given by

$$\widehat{CE_p(f)}^p = \frac{1}{n}\sum_{l=1}^{n}\left[\left|\left| \widehat{E[\mathbf{y}|f(\mathbf{x})]}\Big|_{\mathbf{x}^{(l)}} - f(\mathbf{x}^{(l)}) \right|\right|_p^p \right]$$

- For the estimator of conditional expectation, $\widehat{E[\mathbf{y}|f(\mathbf{x})]}$, we need to estimate the distribution of $(y_j, f_1(\mathbf{x}), \ldots, f_K(\mathbf{x}))$ for all $j \in \{1, \ldots, K\}$. We have used copula for this.
- **N-Copula:** A N-copula(N-dimensional copula) is a function $C : [0,1]^N \to [0,1]$ such that: (1) C is grounded and N-increasing; (2) C has margins $C_i, i \in \{1, \ldots, N\}$ and $C_i(u) = u, u \in [0,1]$.
- **Sklar's Observation:** Observation by Sklar suggests that univariate marginals and multivariate dependence structure can be modeled separately and the information of dependence can be represented by a Copula. So, a copula encapsulates all the necessary information to measure dependence and remains invariant under any nonlinear, strictly increasing transformations of the marginal variables.

Sklar's Theorem

Consider N dimensional CDF F with marginals $F_1, \ldots, F_N$. Then there exists a N-copula, C, such that

$$F(x_1, \ldots, x_N) = C(F_1(x_1), \ldots, F_N(x_N))$$

As copula is extremely useful when it comes to model dependencies, we take a semiparametric approach to estimate calibration error where copula is modeled parametrically and the marginal distributions are modeled non-parametrically.

- By Sklar's theorem, the joint CDF of $(y_1, \ldots, y_K, f_1(\mathbf{x}), \ldots, f_K(\mathbf{x}))$ can be written as $C(\mathfrak{F}_1(y_1), \ldots, \mathfrak{F}_K(y_K), F_1(f_1(\mathbf{x})), \ldots, F_K(f_K(\mathbf{x})))$.
- We consider the $(K+1)$ margin of C on variables $(y_j, f_1(\mathbf{x}), \ldots, f_K(\mathbf{x}))$ for all $j \in \{1, \ldots, K\}$. This margin will naturally inherit the information of label interdependencies from the $2K$-copula C.
- $E[y_j|f(\mathbf{x})]$ can be written as

$$E[y_j|f(\mathbf{x})] = \int_{-\infty}^{+\infty} y_j g_{y_j}(y_j|f(\mathbf{x}))dy_j = E\left[y_j \frac{c(\mathfrak{F}_j(y_j), F_1(f_1(\mathbf{x})), \ldots, F_K(f_K(\mathbf{x})))}{c_f(F_1(f_1(\mathbf{x})), \ldots, F_K(f_K(\mathbf{x})))} \right]$$

- Need to find $\hat{\mathfrak{F}}_j, \hat{F}_1, \ldots, \hat{F}_K$ and estimation of copula densities c and c_f in order to find $\widehat{E[y_j|f(\mathbf{x})]}$.

- y_j's are discrete random variables but Sklar's theorem is not applicable for discrete variables. So, we construct a continuous CDF of y_j by leveraging kernel trick and we take empirical CDF as the estimator of CDF for $x_j, j \in \{1, \ldots, K\}$. They are defined as follows:

$$\hat{\mathfrak{F}}_j(y_j) = \begin{cases} 0 & \text{if } y_j < -0.5 \\ \hat{p}_j(0)(y_j + 0.5) & \text{if } -0.5 \le y_j \le 0.5 \\ \hat{p}_j(0) + \hat{p}_j(1)(y_j - 0.5) & \text{if } 0.5 \le y_j \le 1.5 \\ 1 & \text{if } y_j > 1.5 \end{cases} \quad \textit{and} \quad \hat{F}_j(f_j(\mathbf{x})) = \frac{\sum_{i=1}^{n} \mathbb{1}(f_j(\mathbf{x}) \le f_j(\mathbf{x}^{(i)}))}{n}$$

- To define the copula in a parametric manner, we could use the Maximum Likelihood Estimation (MLE). However, directly estimating θ using MLE is challenging because $\mathfrak{F}_j$ and F_j for $j \in \{1, \ldots, K\}$ are not empirically tractable. To address this issue, we instead employ the Maximum Pseudo-Likelihood Estimation (MPLE).
- The estimator of $E[y_j|f(\mathbf{x})]$ can be written as

$$\widehat{E[y_j|f(\mathbf{x})]} = \frac{\sum_{i=1}^{n} y_j^{(i)} c\left(\hat{\mathfrak{F}}_j(y_j^{(i)}), \hat{F}_1(f_1(\mathbf{x})), \ldots, \hat{F}_K(f_K(\mathbf{x})); \hat{\theta}_j\right)}{\sum_{i=1}^{n} c\left(\hat{\mathfrak{F}}_j(y_j^{(i)}), \hat{F}_1(f_1(\mathbf{x})), \ldots, \hat{F}_K(f_K(\mathbf{x})); \hat{\theta}_j\right)}$$

- The estimator of $CE_p(f)$ is

$$\widehat{CE_p(f)}^p = \frac{1}{n}\sum_{l=1}^{n}\left[\left|\left| \left[\frac{\sum_{i=1}^{n} y_j^{(i)} c\left(\hat{\mathfrak{F}}_j(y_j^{(i)}), \hat{F}_1(f_1(\mathbf{x})), \ldots, \hat{F}_K(f_K(\mathbf{x})); \hat{\theta}_j\right)}{\sum_{i=1}^{n} c\left(\hat{\mathfrak{F}}_j(y_j^{(i)}), \hat{F}_1(f_1(\mathbf{x})), \ldots, \hat{F}_K(f_K(\mathbf{x})); \hat{\theta}_j\right)}\Big|_{\mathbf{x}^{(l)}} \right]_j - f(\mathbf{x}^{(l)}) \right|\right|_p^p \right] \tag{1}$$

Results On Consistency, Differentiability and Biasness

- The estimator of the conditional expectation, $\widehat{E[y_j|f(\mathbf{x})]}$, is unbiased and consistent.
- Hence, the estimator of L_p calibration error, $\widehat{CE_p(f)}^p$, is consistent for all p.
- $\widehat{CE_p(f)}^p$ is unbiased for $p = 1$.
- In our experiments we have chosen Gaussian copula density function as a choice for our parametric copula. So, $CE_p(f)^p$ is inherently differentiable. That means it can be used in training time for calibration regularized risk minimization.

Calibration Regularized Risk Minimization

- Most of the loss functions that are used to train MLC tasks are designed to achieve consistency in terms of Bayesian risk optimization. They are not guaranteed to achieve the same for calibration.
- Our goal is to develop a model f that is both accurate and well-calibrated. To achieve this, we optimize using $CE_p(f)^p$ alongside a loss function.
- We bound $CE_p(f)^p$ by B and set up the optimization problem:

$$f = \arg\min_{f \in \mathcal{H}} \text{Risk}(f) \quad \text{s.t.} \quad CE_p(f)^p < B$$

with the corresponding Lagrangian:

$$f = \arg\min_{f \in \mathcal{H}} \Big(\text{Risk}(f) + \beta \cdot CE_p(f)^p \Big)$$

where $\mathcal{H}$ represents the class of all possible models.

Experiments and Results

- We have demonstrated that adding calibration error estimation as an auxiliary loss to cross-entropy loss during neural network training maintains accuracy while improving model calibration. We call this approach Calibrated Multi-Label Classification by Copula (CMLCC).
- For our experiments we have used ResNet50 along with the famous PASCAL-VOC 2012 dataset and Kaggle Image [1] dataset for multi-label classification.

Dataset	CE	FL	CMLCC
PASCAL-VOC	0.03819 $\pm$0.0048	0.2034 $\pm$0.036	**0.022** $\pm$0.0005
Kaggle Image	0.046 $\pm$0.0004	0.1484 $\pm$0.006	**0.012** $\pm$0.0002

Table 1. ECE

Dataset	CE	FL	CMLCC
PASCAL-VOC	0.0711	0.0762	**0.0701**
Kaggle Image	0.0379	0.0402	**0.0378**

Table 2. Hamming Loss

References

[1] Multi label image classification dataset, `https://www.kaggle.com/dsv/3943866`.

[2] Tianshui Chen, Weihang Wang, Tao Pu, Jinghui Qin, Zhijing Yang, Jie Liu, and Liang Lin.
Dynamic correlation learning and regularization for multi-label confidence calibration.
IEEE Transactions on Image Processing, 2024.

[3] Jiacheng Cheng and Nuno Vasconcelos.
Towards calibrated multi-label deep neural networks.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27589–27599, 2024.

[4] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger.
On calibration of modern neural networks.
In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[5] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht.
Obtaining well calibrated probabilities using bayesian binning.
In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.